A Local and Global Feature Disentangled Network: Toward Classification of Benign-malignant Thyroid Nodules from Ultrasound Image

Shi-Xuan Zhao, Yang Chen, Kai-Fu Yang, Yan Luo, Bu-Yun Ma, and Yong-Jie Li

***Abstract*— Thyroid nodules are one of the most common nodular lesions. The incidence of thyroid cancer has increased rapidly in the past three decades and is one of the cancers with the highest incidence. As a non-invasive imaging modality, ultrasonography can identify benign and malignant thyroid nodules, and it can be used for large-scale screening. In this study, inspired by the domain knowledge of sonographers when diagnosing ultrasound images, a local and global feature disentangled network is proposed to classify benign and malignant thyroid nodules. This model imitates the dual-pathway structure of human vision and establishes a new feature extraction method to improve the recognition performance of nodules. We use the tissue-anatomy disentangled (TAD) block to connect the dual pathways, which decouples the clues of local and global features based on the self-attention mechanism. To verify the effectiveness of the model, we construct a large-scale dataset and conduct experiments. The results show that our method achieves an accuracy of 89.33%, which has the potential to be used in the clinical practice of doctors, including early cancer screening procedures in remote or resource-poor areas.**

***Index Terms*— ultrasound image, thyroid nodule, attention mechanism, classification, deep neural network**

I. Introduction

Thyroid nodules are a prevalent nodular lesion. About 4–5% of adults in the United States have identifiable thyroid nodules, and statistics indicate that the incidence is still rising [1], [2]. In addition, new cases of malignant nodules due to thyroid cancer have increased 2.41-fold in the last 35 years, from 5.8 cases per 100,000 men and women in 1992 to 14.0 cases per 100,000 men and women in 2016 [3]. Although the incidence of thyroid nodules is high, the proportion of malignant nodules is less than 10% [4], [5]. Typically, only benign nodules need to be reviewed regularly without surgical treatment unless the nodules are significantly enlarged or the patient has symptoms such as difficulty swallowing or voice changes. For cancer patients to receive the proper treatment and patients with benign nodules to avoid unnecessary treatment or surgery, it is vital to accurately distinguish between benign and malignant nodules.

There are two commonly used methods for diagnosing benign and malignant thyroid nodules: noninvasive ultrasonography [6] and invasive fine-needle aspiration biopsy (FNAB) [7]. FNAB is the gold standard for nodule diagnosis, but the use of FNAB for large-scale screening subjects the patients to trauma and incurs additional costs. In contract, ultrasound imaging is fast, low-cost, and does not use radiation, plus it can obtain high-resolution images without causing damage to the superficial organs of patients. It is suitable for checking thyroid gland health in people of all ages and is now one of the most commonly used examination methods.

In 2009, Horvath et al. [8] and Park et al. [9] independently proposed Thyroid Imaging Reporting and Data System (TI-RADS, as shown in Table I) inspired by breast imaging reporting and data system (BI-RADS). TI-RADS aims to assess thyroid nodules more objectively and avoid unnecessary invasive examinations. Based on thyroid ultrasound images, features such as shape, orientation, edge, calcification, and echogenicity are used as ultrasound descriptors to stratify the risk of thyroid nodule malignancies. Subsequent research [10], [11] has also provided improvements to some grading schemes and indicated that TI-RADS diagnosis has a high discriminative power, which is mainly reflected in the sensitivity of diagnosis. However, there are still several obstacles that have limited the diagnostic effect of TI-RADS. First, due to insufficient awareness of the features of malignant and benign nodules, experts still have disputes concerning the interpretation and recommendation of nodules in clinical practice. Second, the judgment of sonographers is subjective and relies heavily on a large amount of experience, which increases the difficulty of large-scale screening. Therefore, automatic and accurate classification of thyroid nodules based

This work was supported by Key Area R&D Program of Guangdong Province (2018B030338001); National Natural Science Foundation of China (61806041, 62076055); Department of Science and Technology of Sichuan Province (2021YJ0245). Shi-Xuan Zhao and Yang Chen have contributed equally to this work. (Corresponding authors: Bu-Yun Ma and Yong-Jie Li.)

Shi-Xuan Zhao, Kai-Fu Yang, and Yong-Jie Li are with the Radiation Oncology Key Laboratory of Sichuan Province, MOE Key Lab for Neuroinformation, University of Electronic Science and Technology of China, Chengdu, China. Email: liyj@uestc.edu.cn.

Yan Luo, Bu-Yun Ma are with the Department of Ultrasound, West China Hospital, Sichuan University, Chengdu, China. Email: mabuyundoctor@163.com.

Yang Chen is with the Department of Ultrasound and the West China Biomedical Big Data Center, West China Hospital, Sichuan University, Chengdu, China.

TABLE I
TI-RADS SCORES AND DESCRIPTIONS

Scores	Descriptions
1	Normal gland
2	Benign nodule
3	Highly probable benign nodule
4A	Low suspicion for malignacy
4B	High suspicion for malignacy
5	More than two criteria of high suspicion for malignacy

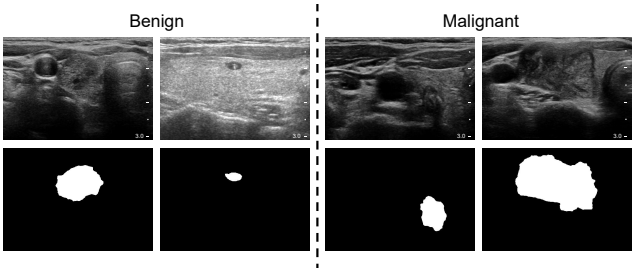

Fig. 1. The first row contains thyroid ultrasound images (the rightmost of the image is the scale in cm.), and the second row contains rough binary segmentation templates that level set algorithm converges the hint point of the sonographers.

on ultrasound images is a critical topic.

We first explore this research question from the perspective of the doctor. As shown in Fig. 1, the high complexity of the thyroid nodule location (containing many tissues such as the trachea, arteries, and blood vessels) and the significant differences of nodule shape and size are inherent characteristics of the ultrasound images of the thyroid, which increases the difficulty of model training. We summarized the features observed by human experts (i.e., imaging manifestation) during diagnosis: the local features (internal echo intensity, texture, definition of boundary, aspect ratio) and the global features (location of the thyroid, surrounding echo, relative size). It can be seen that the local features rely on the location of the tissues, whereas the global features depend on the perception of the overall anatomy structure, and these two types of features complement each other. Effective use of these two types of features is beneficial to accomplish diagnosis [12], [13].

An emerging technique called radiomics aims to quantitatively analyze the characteristics of medical image data, making it possible to obtain disease characteristics unrecognizable to the naked eye [14]. Radiomics has also been applied to the benign and malignant classification tasks of thyroid nodules, which mainly fall into two broad categories: the traditional methods [15], [16] and the deep learning methods [17], [18].

The traditional radiomics methods use the manual design feature extraction method combined with feature selection and classifiers for diagnosis. However, there are two essential obstacles with traditional methods: (1) They rely on excellent tissue's contours to ensure the stability of feature extraction, which increases a significant labor cost. The subjectivity of the contours by doctors will cause the bias of features to affect the generalization ability of the classification model. (2) It is challenging to design effective feature extraction algorithms. In the current radiomics research, the mainstream feature extraction algorithms are summarized in [19], which are still focused on the lesion local features, lacking attention to the global features.

The deep learning methods for thyroid nodule diagnosis are based on the deep neural network (DNN). DNN can be regarded as an end-to-end mapping from image to class labels, which has functions such as feature extraction, feature selection, and classification. Although DNN has a powerful ability in mapping complicated nonlinear input-output relationships, the introduction of medical domain knowledge into deep learning methods can help better break through the difficulties of medical tasks [20].

As mentioned earlier, only appropriate combination of the local features based on the organization and the global features based on the anatomical structure can better complete the classification of thyroid nodules. Some studies have attempted to extract global and local features in specific ways. He et al. [21] extracted local features from thyroid nodules delineated by doctors with a self-organizing map (SOM) method on ultrasound images and represented the elasticity information as global features from a more extensive range around the nodules, which achieved good performance in nodule classification. Similarly, Ding et al. [22] achieved significant performance in the classification of pulmonary nodules by extracting the feature like the overall appearance (OA), heterogeneity in voxel values (HVV), and heterogeneity in shapes (HS) of pulmonary nodules to make predictions. Although these studies have made efforts to extract local and global features, they are limited to obtaining different fields of view by changing the input of the convolutional neural network (CNN) to extract diversified features. However, due that convolutional operations only deal with the local neighborhood for image data, establishing long-range dependencies (used to obtain global features) needs deep stacks of convolution operations to increase the receptive fields [23], which is inefficient and difficult to optimize [24]. This study hopes to establish a more suitable way to decouple the local and global features in the feature space to complete the classification task.

The human visual system (HVS) is a complex perception system formed after hundreds of millions of years of evolution, which can deal with complex environments and is accurate, efficient, and controllable. Neurophysiological and neuropsychological research has confirmed the existence of the ventral system and the dorsal system, which are functionally different [25]. In the HVS, the ventral system has the function of target recognition, the so-called "What" pathway, and the dorsal system has the ability of spatial positioning, the so-called "Where" pathway [26]. More and more evidence has shown that the two pathways have functional interactions [27].

This study is inspired by the HVS and proposes a local and global feature disentangled network (LoGo-Net) that aims to establish a novel feature expression model to classify benign and malignant thyroid nodules using ultrasound images. We use the self-attention mechanism [28], [29] to design a tissue-anatomy disentangle (TAD) block that bridges the "What" and "Where" pathways. It separates the tissue components used to extract local features and the anatomy components used to extract global features from the ultrasound image. The

contributions of this paper are threefold:

- A novel local and global feature representation learning method is established using medical domain knowledge, which achieves state-of-the-art performance on the task of thyroid nodule classification.
- We propose a feature disentangled block that connects the "What" pathway and the "Where" pathway as a bridge of information interaction, consequently establishing an original classification and segmentation multi-task learning network.
- This study explores the use of artificial intelligence (AI) to assist doctors, providing a possible joint diagnosis mode, which illustrates its potential for clinical application by improving the efficiency of large-scale disease screening and the diagnosis capabilities.

II. Related works

A. Traditional methods for the diagnosis of thyroid nodules

Typically, image-based traditional methods for thyroid nodule diagnosis include three parts: feature extraction, feature selection, and classifiers. Among them, feature selection and classifiers are relatively mature, and feature extraction is the core problem, i.e., how to design manually designed features to express the echo state of thyroid nodules. Nam et al. [15] proposed a histogram-based analysis to find the difference between benign and malignant thyroid nodules from ultrasound images, but this approach did not improve the diagnostic effect compared to experts. Raghavendra et al. [16] extracted spatial gray level dependence features (SGLDF) and fractal textures to describe the structure of the thyroid and then used marginal Fisher analysis (MFA) and support vector machine (SVM) to perform feature selection and classification, respectively, and achieved good performance. Song et al. [30] used a gray level co-occurrence matrix (GLCM) to extract texture features and six different statistical models for classification. It was found that the combination of GLCM and logistic model can help distinguish benign and malignant thyroid nodules. Nugroho et al. [31] constructed feature extraction through histogram, GLCM, and gray level run length matrix (GLRLM) and used multilayer perceptron (MLR) to distinguish cystic nodules from solid nodules.

B. Deep learning methods for the diagnosis of thyroid nodules

Deep neural networks (DNNs) have made breakthroughs in many fields, including medical image analysis. They can complete medical image classification [32] and segmentation [33] and also can build a multi-task learning model [34]. There are also some studies on the classification of thyroid nodules that use deep learning. Chi et al. [17] divided thyroid nodules into benign and malignant by TI-RADS, extracted the image features by GoogLeNet, which uses transfer learning for initialization, performed classification using a random forest classifier, and obtained high-grade classification results. Ma et al. [18] used the histopathological report obtained by surgery or FNAB as the ground truth and proposed two CNNs with different depths to classify thyroid nodules. Liu et al. [35] fused the CNN features and the handcrafted features to build a hybrid feature space. Wang et al. [36] proposed an attention-based method to automatically integrate multiple ultrasound image features and predict the benign and malignant nodules from different views. Although deep learning has a potent feature generalization performance, usually the features extracted by CNN contain various image components, which introduces serious problems such as model overfitting and low robustness.

Specifically, on the task of thyroid nodule classification, the range that the model needs to perceive should cover the overall anatomy structure in addition to the internal region of the nodule. But for DNN, it is difficult to simultaneously extract the precise local nodule features and global features. To deal with this problem, Liu et al. [37] proposed three branches to extract multi-view features for diagnosis, including a basic branch with only images of nodules as input, a context branch with a surrounding region of nodules as input, and a margin branch with the margin regions around nodules as input. In this research, we hope that the model can have a wider range of perceptions to acquire more effective global features. Hence, some studies have begun to investigate feature disentangled methods, as summarized below.

C. Feature disentangled methods in medical image analysis

Feature disentangled methods aim to use the differences of homologous features entangled with each other to separate independent features containing different components, which can better serve downstream tasks. Pei et al. [38] decomposed the source domain and target domain features into domain-invariant features (DIFs) and domain-specific features (DSFs), delivering promising performance in cross-modal cardiac segmentation tasks. In particular, they introduced a self-attention module into encoders to enhance the representation of DIFs. Chartsias et al. [39] learned the decomposition mode of anatomical and imaging factors based on the magnetic resonance imaging (MRI) of the abdomen, where anatomical factors can be used to align anatomical structures of different modalities, and imaging factors can be used to extract signal strength characteristics of different modalities. Ment et al. [40] proposed mutual information-based disentangled networks (MIDNet), decoupled categorical features and domain features based on mutual information minimization, and finally demonstrated the strong potential for fetal ultrasound imaging applications. Tang et al. [41] proposed the hypothesis that disease regions are superimposed upon or replace the pixels of normal tissue on chest X-ray (CXR) images, and the disease regions are separated from abnormal CXR through adversarial learning.

In this study, separating entangled local and global features from thyroid ultrasound images is essential to improving diagnostic performance. It is similar to previous studies in that the feature extraction process of DNN couples features of multiple components. However, since their tasks are different from ours, we have proposed a new disentangled method specifically for thyroid ultrasound images.

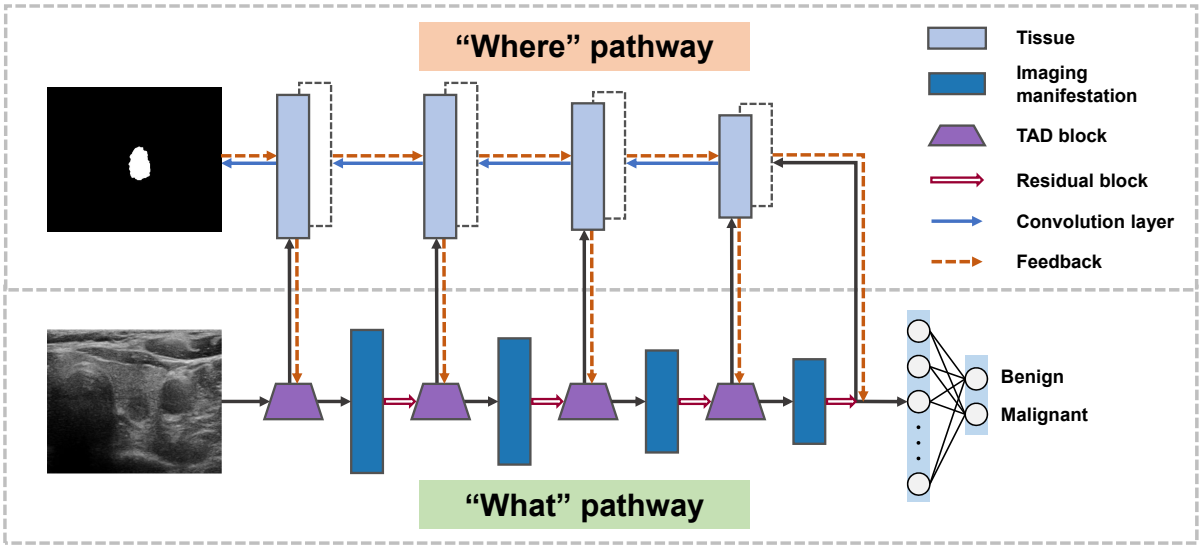

Fig. 2. The overall architecture of the LoGo-Net.

D. Multi-task learning in medical image analysis

Multi-task learning (MTL) is a learning paradigm of machine learning, which can use the correlation between multiple sub-tasks to improve performance. In recent years, MTL has also been widely used in medical imaging tasks. Moeskops et al. [42] attempted to use a single CNN to implement multiple segmentation tasks in multi-modal scenarios, proving the high capacity of the deep learning architecture. At the same time, some studies are dedicated to combining different types of tasks so that the model can dig out common ingredients in diversified tasks to improve the performance of all subtasks. Zhou et al. [43] jointly trained the model on the breast tumor segmentation and classification tasks and obtained better results than single-task learning. Foo et al. [44] proposed a semi-supervised learning process to obtain the diabetic retinopathy region segmentation mask and then combined with the classification network to complete more fine-grained grading of the diabetic retinopathy severity level, which proves the effectiveness of MTL by CNNs. According to fetal head ultrasound image characteristics, Lin et al. [45] designed a MTL framework for standard plane inspection and quality evaluation. It achieves promising performance at nearly real-time processing speed and reduces measurement errors caused by improper scanning. From another perspective, Chen et al. [46] focused on combining multiple tasks, exploring the performance of joint training and alternating training strategies, and using a small amount of labeled data to obtain better performance in brain tumor and white matter hyperintensities segmentation.

This study starts from medical domain knowledge and proposes the TAD block as a bridge between the thyroid nodule segmentation and classification tasks, providing a novel design for multi-task learning.

III. Method

A. Overall architecture

The proposed method is inspired by the human visual system to establish a novel feature representation model for thyroid nodule classification, as shown in Fig. 2. We design a tissue-anatomy disentangle (TAD) block to connect the two pathways, which decouples the features into tissue information containing local feature clues and anatomy information containing global feature clues. The two pathways promote simultaneous optimization through the multi-task learning mode, and more accurate classification can be obtained. The proposed model only needs segmentation masks during training but not during testing, i.e., no further burden is required for segmentation when our trained model is applied in practice.

B. Tissue-anatomy disentangled block

The calculation mode of a non-local (NL) block [47] works as self-attention to select and focus a small amount of important information from an extensive region, ignoring most worthless information, which is very consistent with the selective attention mechanism of human vision. Specifically, a NL block calculates the similarity between two pixel sets (referred to as $query$ and key), which is capable of capturing the long-distance dependence of features. Then, the weight coefficient matrix obtained is used to weight the current features (referred to as $value$), as follows:

$$\mathbf{y}_i = \omega(\mathbf{x}_i, \mathbf{x}_j) \times \mathbf{v}_j = \mathcal{S}(\mathbf{q}_i^T \mathbf{k}_j) \times \mathbf{v}_j, \tag{1}$$

where $\mathbf{x}$ denotes the input feature, $\mathbf{y}$ denotes the output feature of the non-local block, and i and j denote the feature position indexes, so $\omega(\mathbf{x}_i, \mathbf{x}_j)$ represents the function of measuring the embedded similarity of $\mathbf{x}_i$ and $\mathbf{x}_j$. When $\omega(\mathbf{x}_i, \mathbf{x}_j)$ is instantiated with embedded Gaussian function, it becomes a softmax computation along the dimension of $\mathbf{x}_j$, expressed as $\mathcal{S}(\mathbf{q}_i^T \mathbf{k}_j)$. $\mathbf{q}_i = W_q \mathbf{x}_i$, $\mathbf{k}_j = W_k \mathbf{x}_j$, and $\mathbf{v}_j = W_v \mathbf{x}_j$ respectively represent $query$, key, and $value$, where W_q, W_k, and W_v are weight matrices to be learned.

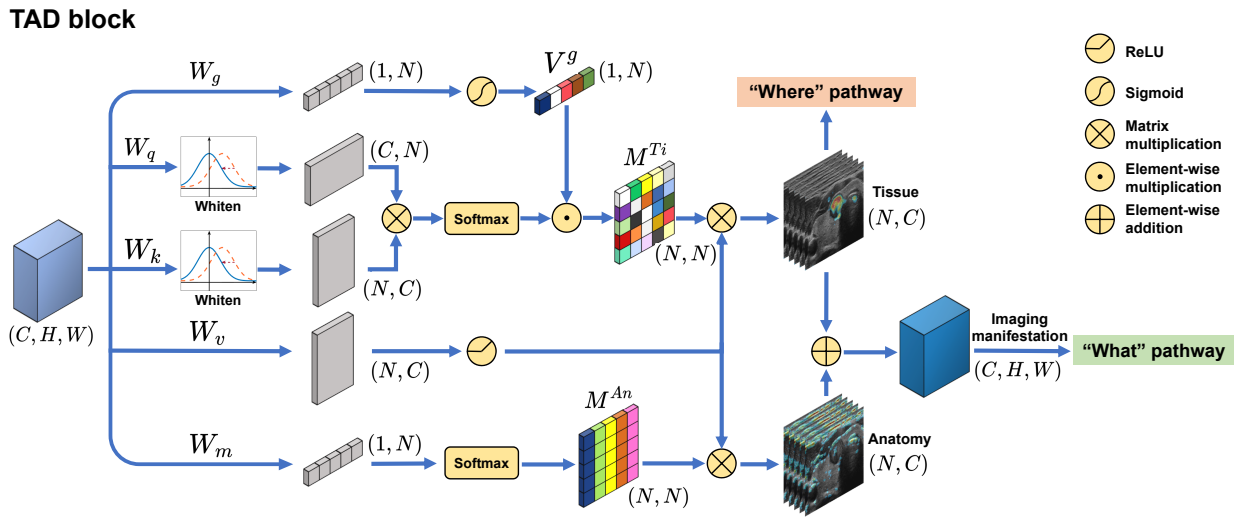

Fig. 3. The details of the proposed tissue-anatomy disentangled (TAD) module. M^{Ti} represents the attention maps of tissue, and M^{An} represents the attention maps of anatomy.

Yin et al. [48] found there are two components encoded in the attention maps obtained by a NL block encoding: a pairwise term containing the correlation between pixels and a unary term inferring the saliency of each pixel. However, the gradients of both components are determined by the value of the other term, so when one term is optimized, the weight of another term will be close to zero. Hence, the coupling of the two components hinders their optimization because of the gradient vanishing. They designed the disentangled non-local (DNL) block to separate these two pairwise and unary terms [48], expressed as:

$$\omega^D(\mathbf{x}_i, \mathbf{x}_j) = \mathcal{S}(\underbrace{(\mathbf{q}_i - \boldsymbol{\mu}_q)^T(\mathbf{k}_j - \boldsymbol{\mu}_k)}_{\text{pairwise}} + \underbrace{\boldsymbol{\mu}_q^T \mathbf{k}_j}_{\text{unary}}), \quad (2)$$

where $\boldsymbol{\mu}_q = \frac{1}{|\Omega|}\sum_{i\in\Omega}\mathbf{q}_i$ and $\boldsymbol{\mu}_k = \frac{1}{|\Omega|}\sum_{j\in\Omega}\mathbf{k}_j$, and Ω denotes all pixels of a feature map with a size of $H \times W$ pixels. DNL proves that compared with the attention of a general NL block calculation, pairwise terms are more focused on the attention of within-category, and unary terms are focused on the boundary [48].

Although DNL provides two clues for feature extraction, it does not integrate them with the task precisely. In addition, in this task, the network needs to pay more attention to the local features of the nodules rather than other tissues contained in the image, so it needs to further restrict unary terms. In this study, we propose three concepts, i.e., the imaging manifestation, the tissue, and the anatomy. The imaging manifestation fuses the tissue and anatomy components, which constitute the diagnosis feature representation extracted by the model. The tissue component represents the local feature clues of the image, including the intensity, texture, and shape features of the nodules. In contrast, the anatomy component provides global feature clues for extracting the features of location, context, and relative size of nodules.

Hence, we propose the TAD block to disentangle the tissue and anatomy components from the feature maps, as shown in Fig. 3. Unlike the DNL block, our TAD block can make the attention of pairwise terms focused on the within-category related to nodules for classification so as to avoid model overfitting caused by the introducing of irrelevant tissues. Specifically, for the tissue component, we focus on the attention of within-category regions through the NL block by using the whitening operation on key and $query$, and we obtain the attention maps of various tissues in thyroid images. Then we use the gating unit $V^g = \sigma(\mathbf{g}_j)$ to let the feature maps focus on specific areas in the spatial domain. The nodule location constraint from the "Where" pathway can optimize the parameters in the gating unit, making the model selectively focus on the nodules and other related tissues for classification. For the anatomy component, we set a new weight matrix W_m that is not shared with W_k, which causes the optimization of tissue and anatomy components to be independent of each other. This form allows anatomy components not to be confined to the nodule surrounding but to capture the feature with a larger field of view. Due to the optimization of "What" pathway for classification, anatomy components will focus on perceiving the thyroid scene structure for supplementing the feature cues that tissue components cannot provide.

Naturally, in the process of model optimization, TAD block decouples the feature map into tissue and anatomy components, which is written as follows:

$$\begin{aligned}\mathbf{y}_i &= \omega^G(\mathbf{x}_i, \mathbf{x}_j) \times \rho(\mathbf{v}_j) \\ &= \underbrace{\sigma(\mathbf{g}_j) \cdot \mathcal{S}\left((\mathbf{q}_i - \boldsymbol{\mu}_q)^T(\mathbf{k}_j - \boldsymbol{\mu}_k)\right) \times \rho(\mathbf{v}_j)}_{\text{Tissue}} + \underbrace{\mathcal{S}(\mathbf{m}_j) \times \rho(\mathbf{v}_j)}_{\text{Anatomy}},\end{aligned} \quad (3)$$

where $\rho(\cdot)$ represents the ReLU activation function, and $\sigma(\cdot)$ represents the sigmoid activation function. $\mathbf{g}_j = W_g\mathbf{x}_j$ and $\mathbf{m}_j = W_m\mathbf{x}_j$, where W_g and W_m are two weight matrices to be learned.

C. "What" and "Where" pathways

The "What" pathway realizes the function of benign and malignant thyroid nodule classification, which is the trunk

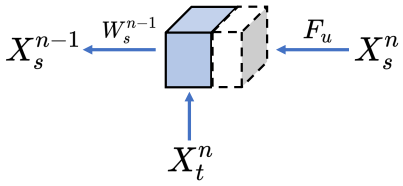

Fig. 4. The details of the decoding unit of the "Where" pathway.

of the proposed framework shown in Fig. 2. In contrast, the "Where" pathway realizes the function of thyroid nodule segmentation, which is a branch of the model.

Regarding the "What" pathway, the backbone of the proposed classification model is composed of residual blocks. Each block contains 3×3 convolution layers to learn feature representation and downsampling with 2×2 to accomplish feature information integration. Before calculating each residual block, we use the TAD block to decouple the feature map into tissue components and anatomy components containing local and global feature clues, respectively. The fusion of tissue and anatomy by the operation of imaging manifestation is sent to the feature extractor so that the model can obtain more abundant features. Furthermore, the tissue components will be fed to the "Where" pathway for information exchange to constrain it and obtain the specific location of the nodule.

Regarding the "Where" pathway, the proposed segmentation model is inspired by the encoder-decoder structure of the U-Net [49], which encodes the high-level features to retain correct semantic information and the low-level features to retain precise details. Our method regards the high-level semantic information of the classification model as the original coarse feature of decoding, continuously integrates various scales of tissue terms through skip connections to recover more precise edge details, and finally obtains the segmentation prediction of nodules. The decoder unit is shown in Fig. 4, which is calculated as:

$$X_s^{n-1} = W_s^{n-1}[F_u(X_s^n), X_t^n], \tag{4}$$

where X_s^n and X_s^{n-1} denote the high-level and low-level feature maps, respectively, X_t^n represents the tissue components, $F_u(\cdot)$ is the upsampling function with a bilinear interpolation calculation, $[\cdot]$ denotes the concatenation, and W_s represents the convolution calculation.

The "What" and "Where" pathways complete end-to-end training simultaneously in the form of multi-task learning and jointly promote convergence. We use the "What" pathway to extract useful discriminant features by optimizing the classification model so that the class discriminant region based on it will benefit the learning of the "Where" path. By optimizing the "Where" pathway, we can obtain the features that reveal the location of the nodule so that we can focus more on extracting the features of the nodule area, which promotes the learning of the "What" pathway.

The loss function $\mathcal{L}$ of the proposed LoGo-Net contains two main parts, i.e., classification loss functions $\mathcal{L}_{cls}$ and $\mathcal{L}_{seg}$. Because of the existence of tiny nodules, the ultrasound images contain highly unbalanced segmentation data, and the nodules are usually much smaller compared with the background. In order to deal with this problem, we use the hybrid loss for segmentation loss. Then the total loss function $\mathcal{L}$ can be written as:

$$\begin{aligned}\mathcal{L} =& \mathcal{L}_{cls} + \lambda \mathcal{L}_{seg} = \mathcal{L}_{cls} + \lambda(\mathcal{L}_{bce} + \mathcal{L}_{iou}) \\ =& -\frac{1}{N}\sum_{b=1}^{N}(y_b \log \hat{y_b} + (1-y_b)\log(1-\hat{y_b})) \\ & -\frac{\lambda}{N}\sum_{b=1}^{N}\left(Y_b \log \hat{Y_b} + (1-Y_b)\log(1-\hat{Y_b})\right) \\ & + \lambda(1 - \frac{\boldsymbol{Y}\cdot\hat{\boldsymbol{Y}}}{\boldsymbol{Y}+\hat{\boldsymbol{Y}}-\boldsymbol{Y}\cdot\hat{\boldsymbol{Y}}}),\end{aligned} \tag{5}$$

where $\mathcal{L}_{bce}$ represents the binary cross-entropy (BCE) loss, $\mathcal{L}_{iou}$ represents the intersection over union (IOU) loss, y denotes the classification label, and $\hat{y}$ denotes the predicted label. $\boldsymbol{Y} = \{Y_1, Y_2, \cdots, Y_N\}$ refers to the segmentation ground truths, and $\hat{\boldsymbol{Y}}$ represents the predicted probabilities. N indicates the batch size. λ is a hyper-parameter that is empirically set to be 0.5 in the experiments.

D. Evaluation metrics and statistical analysis

To evaluate the performance of classification, we adopted some commonly used metrics in this study, including accuracy (ACC), sensitivity (SEN), specificity (SPC), positive predictive value (PPV), and negative predictive value (NPV), which are computed as [50]:

$$\begin{aligned} ACC &= \frac{TP+TN}{TP+TN+FP+FN}, \\ SEN &= \frac{TP}{TP+FN}, \quad SPC = \frac{TN}{TN+FP}, \\ PPV &= \frac{TP}{TP+FP}, \quad NPV = \frac{TN}{TN+FN}, \end{aligned} \tag{6}$$

where TP, TN, FP, and FN denote the numbers of true positives, true negatives, false positives, and false negatives, respectively. In addition, we use the area under the receiver operating characteristic curve (AUC) to measure the stand-alone performance of the AI system. The 95% confidence intervals (CIs) of accuracy, sensitivity, and specificity are calculated by the Clopper-Pearson method. McNemar's test was used to calculate the two-sided P value for accuracy, sensitivity and specificity between models and human experts. The DeLong test was used to calculated the P value for AUC between models. Statistical significance was defined as a P value <0.05.

IV. Experiment and Results

A. Materials

1) Subjects and data collection: Although there are some related studies on the classification task of benign and malignant thyroid nodules, there is no large-scale, rigorously labeled public dataset with a standardized collection method available for use. In this work, we built a large-scale dataset based on the thyroid nodule ultrasound images collected from West China Hospital (denoted as the THY-BM dataset). This dataset includes 21597 patients, each of which contains one image.

TABLE II
DATA DISTRIBUTION OF THE THY-BM DATASET

	Benign	Malignant	Total
Pathology reports	5931	14057	19988
Ultrasonic reports	1609	0	1609
Total	7540	14057	21597

Then we randomly extracted 2000 samples from the THY-BM dataset to construct a new dataset (denoted as the THY-CHE dataset) to compare our proposed model with human experts.

During ultrasound examination, the ultrasound physicians used a high-frequency linear probe to collect two-dimensional ultrasound images for each patient. More specifically, the ultrasound physicians used the probe to slide and scan the thyroid of the patient continuously, and they saved the image of the largest section of the lesion to the database. Due to the limitations of ultrasound imaging, tiny nodules cannot fully present their image characteristics. Therefore, in this study, we only collected the nodules with a diameter of more than 2mm into study. Dynamic range values, gain values, imaging depth, and frequency are adjusted as needed in order to obtain clear ultrasound image of thyroid. Several brands of ultrasound instruments were used in this retrospective study. In general, the dynamic range value was set from 40 dB to 85 dB, the gain value was set from 50 dB to 70 dB, imaging depth was set from about 2.5 cm to 4 cm, and frequency was from 6 to 12 MHz. The ultrasound instruments included Philips HD 11 XE (Royal Philips Electronics Corporation, Holland Netherlands) with an L12-5 linear probe (Royal Philips Electronics Corporation) with a frequency range of 5 to 12 MHz, Philips ATL HDI 3500 (Royal Philips Electronics Corporation, Holland Netherlands) with an L12-5 linear probe (Royal Philips Electronics Corporation) with a frequency range of 5 to 12 MHz, and SIEMENS Sequoia 512 (Siemens Aktiengesellschaft, Germany Berlin and Munich) with a 15L8W linear probe (Siemens Aktiengesellschaft) with a frequency range of 7 to 12 MHz.

This study and its procedures were approved by the local ethics committees. All methods were performed in accordance with the relevant guidelines and regulations. The entire experiment followed the Helsinki Declaration. Informed consent was not required for this retrospective study (i.e., those discharged or who died). Written informed consent from the involved patients was not required.

2) Data annotation: The THY-BM dataset contains 21597 thyroid nodule ultrasound images, 14057 with malignant nodules, and 7540 with benign nodules. The THY-CHE dataset contains 2000 thyroid nodule ultrasound images, 1288 with malignant nodules, and 712 with benign nodules. Pathological reports were used to determine all malignant nodule labels, and pathological records and ultrasonic reports were used to determine benign nodule labels. The data distribution is summarized in Table II. A total of 1609 benign nodules did not have pathological reports because ultrasound physicians did not perform FNAB or surgical treatment on patients whose ultrasound images manifested benign nodules (i.e., TI-RADS 2). For ultrasound images containing two or more nodules, their labels were annotated as malignant if one nodule was malignant.

Due to the large field of view of thyroid ultrasound, the collected images contain many tissues (e.g., trachea, arteries, and blood vessels) and noise, as shown in Fig. 1, making it difficult to position the nodules accurately without a sketch made by an experienced doctor. Finely delineating nodules is a massive workload for doctors, so we used a scheme to lessen the sketching requirements in this study. In detail, the doctors hinted a few points near the contour of thyroid nodules on each ultrasound image as its initial contour. Then we employed the level set algorithm [51] to adaptively extract its rough binary segmentation template.

3) Implementation: All the models were trained and tested on the Pytorch platform, using servers equipped with NVIDIA RTX 3090 GPUs. During the training process, ultrasound images were rescaled to 224×224 and normalized before being input to the network. The binary segmentation templates were rescaled to 56×56. We used the stochastic gradient descent (SGD) algorithm to optimize the loss function, in which momentum was set to 0.9 and weight decay was set to 1e-5. Models were trained in a total of 100 epochs with a batch size of 16, in which the learning rate was initially set to 0.001 with a decay rate of 0.95 per epoch. To deal with the problem of data imbalance during training, we adopted an over-sampling approach, i.e., all the negative samples were randomly repeated and flipped horizontally to the same number of positive samples in an epoch. All results were based on five-fold cross-validation. Our code can be downloaded at `https://github.com/Phanzsx/LoGo-Net`.

B. Performance of thyroid nodule classification

1) Comparison of different methods: We compared the results of different methods with the two input modes (i.e., the ROI and the full image), as shown in Table III. The ROI mode means the model takes the bounding box containing the nodule area located by the doctor as the input. This mode breaks the original scale of the nodule by resetting the ROI size required by CNN, which leads to the change of some shape and texture features of the nodule. As shown in Table III, with ROI as input, the classification indexes achieve good performance, and the differences of AUCs among various models are slight, indicating that feature extraction within the limited ROI region makes it easy to obtain the feature presentation of the nodule.

The full image mode means that the entire thyroid ultrasound image, containing thyroid nodules and other nearby organs or tissues, is taken as the input to the model. This mode contains the global image information and preserves the invariance of the original scale features. For the existing methods, the overall performances are not as good as that of the ROI mode, as shown in Table III. This indicates that the full image mode covers more image information but increases the difficulty of extracting practical features. In contrast, the proposed method is capable of extracting the tissue containing faithful nodule features as well as the anatomy information containing global structural features through the disentangled

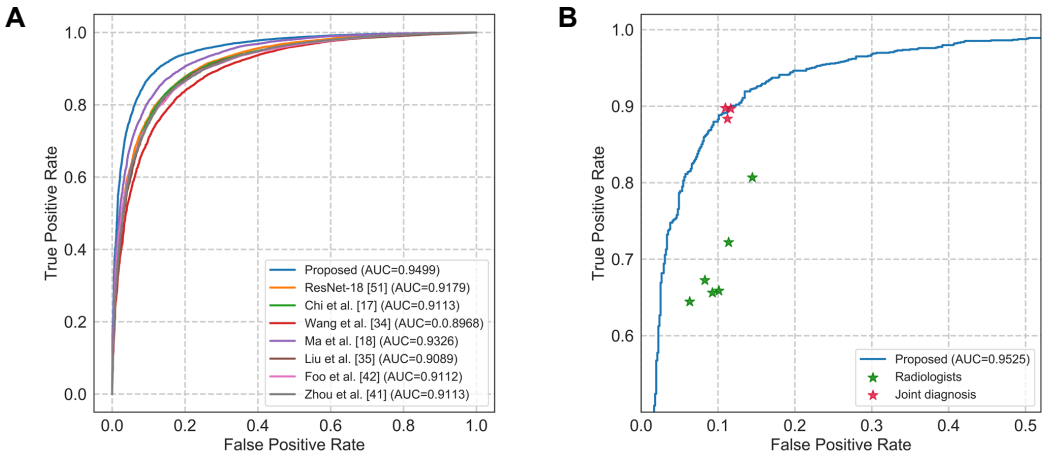

Fig. 5. A: The receiver operating characteristic (ROC) curves of the different networks with full image input mode on the THY-BM dataset. B: Comparison of our LoGo-Net with six human experts provided with only images and three experts assisted with our model's prediction on THY-CHE dataset.

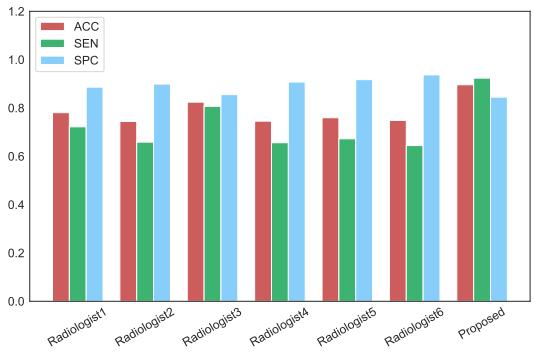

Fig. 6. The bar graph showing the specific diagnosis results of our AI model and human experts.

block in the full image mode, which achieves the best performance compared with the state-of-the-art CNNs, as shown in Table III and Fig. 5A.

We tried a combination of the radiomics method and machine learning classifier (Radiomics-SVM) to classification. The experiment adopt a total of 102 features advocated by [19] (including 9 intensity-based features, 18 morphological features, 24 gray level co-occurrence based features, 16 gray level run length based features, 16 gray level size zone based features, 5 neighboring gray tone difference based features, 14 gray level dependence based features), and the classifier uses the SVM method. The results in Tabel III show that compared with deep learning methods, the performance of Radiomics-SVM is not ideal. On the one hand, it reflects that deep learning has a powerful performance driven by big data and exposes the limitations of manual design features. On the other hand, the traditional radiomics method relies on precise segmentation templates, and its result has the potential of further improvement, but with more workforce.

The training of LoGo-net does not rely on a precise segmentation template, and the rough segmentation by the level set algorithm is only used for optimizing the "Where" path to capture the position of nodules. In order to verify

TABLE III

QUANTITATIVE EVALUATION OF DIFFERENT METHODS IN TWO INPUT MODES FOR THYROID NODULE CLASSIFICATION ON THY-BM DATASET. EVALUATION METRICS ARE DESCRIBED IN SECTION III-D. ° INDICATES THAT THE NODULE SEGMENTATION TEMPLATES NEED TO BE PROVIDED DURING THE TEST. CHM MEANS THAT THE SEGMENTATION TEMPLATE PROVIDED DURING TRAINING HAS EXECUTED A CONVEX HULL OPERATION.

Input	Methods	Results (%)					
		ACC	SEN	SPE	PPV	NPV	AUC
ROI	Inception-V3 [52]	85.91	89.49	79.24	88.94	80.17	92.25
	ResNet-18 [24]	86.02	89.25	80.00	89.27	79.97	92.10
	Res2Net-50 [53]	86.61	90.37	79.60	89.20	81.60	92.70
	DenseNet-121 [54]	86.22	88.97	81.09	89.77	79.78	92.53
	Chi et al. [17]	84.69	86.21	81.86	89.86	76.09	91.13
Whole image	°Radiomics-SVM	73.27	75.42	69.26	82.06	60.18	79.58
	GoogLeNet [55]	84.28	88.41	76.56	87.55	77.99	90.66
	Inception-V3 [52]	83.72	88.47	74.85	86.77	77.69	90.07
	ResNet-18 [24]	85.06	88.08	79.42	88.86	78.14	91.79
	Res2Net-50 [53]	86.76	90.29	80.19	89.47	81.58	92.78
	DenseNet-121 [54]	85.47	89.85	77.31	88.07	80.33	91.88
	Wang et al. [36]	83.07	86.94	75.85	87.03	75.70	89.68
	°Ma et al. [18]	86.35	86.87	85.40	91.73	77.72	93.26
	°Liu et al. [37]	84.00	84.54	83.00	90.26	74.23	90.89
	Foo et al. [44]	84.18	86.20	80.41	89.13	75.76	91.12
	Zhou et al. [43]	84.20	85.20	82.35	90.00	74.90	91.13
	LoGo-Net/CHM	88.75	90.77	85.00	91.86	83.16	**95.00**
	LoGo-Net	**89.33**	**91.36**	**85.53**	**92.17**	**84.16**	94.99

this description, we used the masks processed by the convex hull function (CHM) instead of the segmentation templates obtained by the level set algorithm for model training. CHM is rougher than the level set algorithm, which loses the edge information and retains the position information of the nodule. The result in Table III (i.e., LoGo-Net/CHM) shows that using CHM introduces only a slight decline in various indicators, and there is almost no difference in the AUC of the evaluation model classification performance.

2) Competition and cooperation with human experts: In this experiment, we used the THY-CHE dataset to compare our LoGo-Net with six human experts with extensive clinical experience. It needs to be emphasized that both models and experts can only use ultrasound images to diagnose the patients

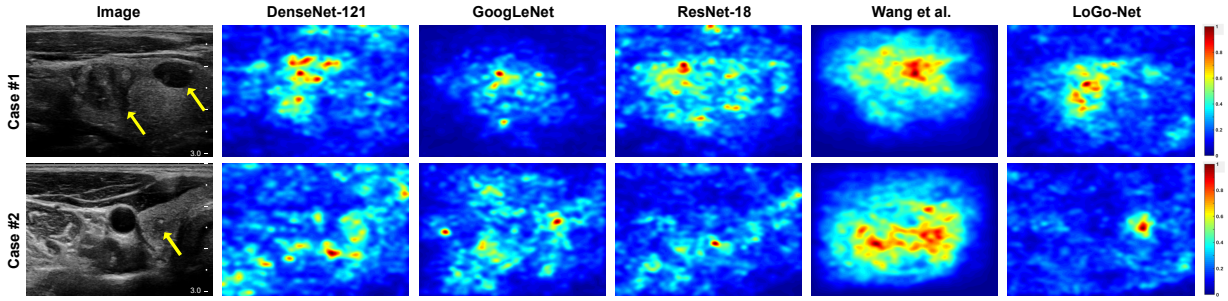

Fig. 7. Visualization of the saliency maps in two cases. The yellow arrows highlight the locations of nodules.

to judge whether the nodules are benign or malignant. THY-CHE contains 2000 patients (712 benign, 1288 malignant) and was not used to train our model. In Fig. 5B, we display the performance of the model as a ROC curve and plot the sensitivity and SPC of the experts. We averaged the results of the six experts, obtaining a 76.70% ACC (95% CI 75.93%, 77.45%), a 69.33% SEN (95% CI 68.29%, 70.36%), and a 90.03% SPC (95% CI 89.09%, 90.91%), and our model achieved an 89.60% ACC (95% CI 88.66%, 90.50%), a 92.35% SEN (95% CI 91.31%, 93.31%), and an 84.48% SPC (95% CI 82.56%, 86.27%). Our model significantly exceeded the average diagnostic level of experts in ACC ($P<0.001$) and sensitivity ($P<0.001$), as shown in Fig. 6.

In order to further explore whether our model can assist doctors in diagnosis, we established a joint diagnosis mode between experts and AI. Three experts participated in the THY-CHE dataset test to diagnose the ultrasound images again and were only provided with the image data the same as the previous. The difference is that in this experiment when the experts think a case is difficult to judge, they can see the prediction of our model for this case (provided in the form of benign and malignant probability). The experts can choose to accept the model's prediction or make a further diagnosis and finally give their judgment. As shown in Fig. 5B, such joint diagnosis clearly improves the accuracy and sensitivity of the diagnosis of the experts and can even reach a level higher than that of the model. So we have reason to believe that with the help of the model, doctors can reach a diagnosis more efficiently and accurately. We think this is a good attempt at how AI can assist doctors.

C. Visualization

1) Analysis of saliency maps: We calculated the saliency maps with the method proposed by [56], which visualizes the contribution of pixels in the image when classifying benign and malignant thyroid nodules. Precisely, the gradient of the loss function of classification can obtain the gradient of the input through a backward pass, and the absolute value of the gradient is the saliency map with the same size as the input. A heat map is generated by normalizing the saliency map, as shown in the second to fifth columns of Fig. 7.

In case #1, the thyroid gland contains two nodules (malignant papillary thyroid carcinoma on the left and benign thyroid nodular goiter on the right). All models focus on malignant nodules, which indicates that networks can capture the characteristics of larger nodules. It is worth mentioning that our LoGo-Net also pays proper attention to the benign nodule on the right side. Although ResNet-18 also pays attention to the benign nodule, it has a broad attention range that is not as concentrated as our method. Furthermore, compared with other networks, the saliency map of our method is more concentrated on the microcalcifications within the malignant nodules, which is more consistent with the experience of experts when distinguishing nodules. In case #2, the entire ultrasound image is more complicated, containing a tiny nodule in the thyroid gland and a large volume of neck muscles and blood vessels. Compared with other networks, the saliency map of our LoGo-Net captures the nodule more accurately, which helps better diagnose the case.

TABLE IV
QUANTITATIVE EVALUATION OF THE DEGREE OF DEVIATION BETWEEN THE SALIENCY MAPS OF DIFFERENT NETWORKS AND THE NODULES.

Methods	DD
GoogLeNet [55]	0.1261
Inception-V3 [52]	0.1872
ResNet-18 [24]	0.1315
Res2Net-50 [53]	0.1570
DenseNet-121 [54]	0.1663
Wang et al. [36]	0.1132
Foo et al. [44]	0.1477
Zhou et al. [43]	0.1226
LoGo-Net	**0.1068**

In order to quantitatively evaluate the matching degree between the model's saliency map and the nodule, we built an evaluation metric named degree of deviation (DD) following the idea of [57], which measures the closeness of the location of the max of the saliency maps to the centroid of nodules, as follows:

$$DD = \frac{1}{N}\sum_{i=1}^{N}\sqrt{\frac{||P_i - C_i||}{H_i^2 + W_i^2}}, \tag{7}$$

where N is the number of images, P_i represents the location of the maximum value pixel in the saliency map, C_i represents the location of the centroid point of the thyroid nodule, H_i and W_i denote the height and width of the image, respectively (the smaller, the better). Compared with other methods, the saliency map of our model has a higher degree of agreement with the nodule location, as shown in Table IV.

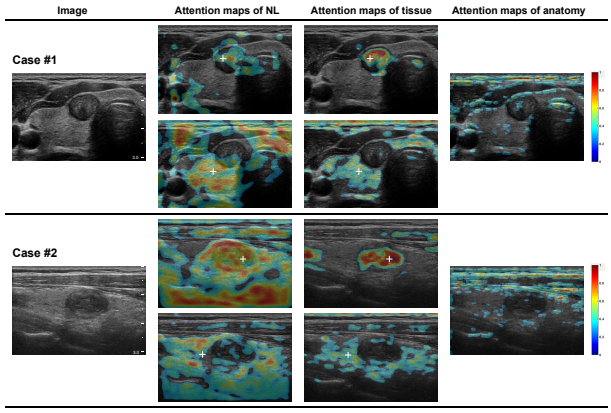

Fig. 8. Visualization of the attention maps for the NL block (column 2) and the TAD block (column 3 and 4). The white plus sign ("+") represents the focus pixel of the non-local calculation.

In summary, the proposed LoGo-Net has excellent recognition ability for nodules with varying morphologies and shapes. The extracted classification discrimination area (i.e., saliency maps) is very consistent with the cognition of experts, which partially enhances the applicability of deep learning.

2) Analysis of attention maps: The attention maps of the NL block and our TAD block provide clues that the model relies on when extracting features. The NL block captures the dependencies between pixels, so its attention maps represent a pixel series potentially related to a single pixel, and we mark a single pixel (i.e., point of interest) in the attention map with a plus sign, as shown in Fig. 8. The attended areas of the NL block are distributed broadly and usually contain several categories of targets. This is because the two terms of the NL block are coupled with each other, and learning one term well to represent the within-category area or boundary area will cause the gradient vanishing of the other term.

In contrast, our proposed TAD decouples the NL block into two terms, i.e., tissue and anatomy, which respectively focus on the pixel relationship in the same category area and the boundary structure information of the body, as shown in the third and fourth columns of Fig. 8. It can be observed that when the point of interest is located inside the nodule and the thyroid gland is outside the nodule, the attention map of tissue provides more precise regional feature clues than that of the NL block. The attention maps of anatomy decoupled by TAD focus on the boundary structure of the body, which captures a more extensive range of feature clues beyond the local texture.

D. Ablation study

To verify the effectiveness of our proposed method, we introduced the dual-pathway structure and the blocks of NL, DNL, and TAD to the baseline model (ResNet-18). Specifically, NL, DNL, and TAD were added to the middle layer of different non-local blocks of the baseline network, and the dual-pathway structure represented the branch that adds the segmentation task to the model (Fig. 2). Table V shows the performance of various combinations (label as A, B, ..., F) and Table VI shows the statistical significance analysis of various combinations' AUC. The performances of introducing the NL block and the dual-pathway structure (B, C, ..., F) separately were improved in all evaluation metrics to varying degrees compared with the baseline network (A), and all the AUC were significantly improved ($P<0.001$). This confirms that both of the long-distance dependency characteristics captured by the NL block and the location constraints from the "Where" pathway can promote classification. In addition, the combination of the feature disentangling blocks (DNL and TAD) with the dual-pathway structure (E, F) had a significant improvement in AUC compared with the combinations excluding DNL and TAD (A, B, ..., D) ($P<0.001$), which proves the effectiveness of our disentangling methods. Although our proposed combination F had no significant improvement on AUC compared with the combination E ($P=0.639$), it clearly improved the ACC and SEN, proving it is a more suitable way to comprehensively consider the function of image disentanglement into tissue components and anatomy components for this task.

TABLE V
QUANTITATIVE EVALUATION OF THE EFFECTIVENESS OF DIFFERENT BLOCKS FOR CLASSIFICATION. THE BASELINE NETWORK A IS RESNET-18.

	NL	DNL	TAD	Dual-pathway	Results (%)			
					ACC	SEN	SPC	AUC
A					85.06	88.08	79.42	91.79
B	✓				87.00	89.08	83.12	93.46
C				✓	87.14	89.59	82.59	93.61
D	✓			✓	86.29	86.16	86.55	93.59
E		✓		✓	87.92	86.88	**89.87**	94.94
F			✓	✓	**89.33**	**91.36**	85.53	**94.99**

TABLE VI
THE STATISTICAL SIGNIFICANCE ANALYSIS OF VARIOUS COMBINATIONS' AUC. THE A, B, ..., F DENOTE THE COMBINATIONS IN TABLE V

	A	B	C	D	E
B	<0.001				
C	<0.001	0.287			
D	<0.001	0.344	0.904		
E	<0.001	<0.001	<0.001	<0.001	
F	<0.001	<0.001	<0.001	<0.001	0.639

TABLE VII
QUANTITATIVE EVALUATION OF DIFFERENT METHODS AIMED AT DATA IMBALANCE.

Methods	Results (%)		
	ACC	SEN	SPE
BCE loss	85.63	91.65	74.41
Focal loss	87.46	90.43	81.91
WBCE loss	88.98	91.10	85.03
Over-sampling + BCE loss	**89.33**	**91.36**	**85.53**

E. Data balance study

To deal with the problem of data imbalance during training, we compared some specific methods' performance, including Focal loss, weighted binary cross-entropy (WBCE) loss, and over-sampling. The results in Table VII show that all methods alleviate the performance degradation caused by data imbalance, and our method (over-sampling + BCE loss) gets the best performance.

TABLE VIII
QUANTITATIVE EVALUATION OF DIFFERENT MALIGNANT/BENIGN RATIOS OF TRAINING DATA.

Malignant/Benign ratio of training data	Results (%)					
	ACC	SEN	SPE	PPV	NPV	AUC
1:9 (total 8378)	88.45	88.55	88.25	93.36	80.53	94.20
2:1 (total 8378)	88.63	89.48	87.04	92.79	81.61	94.68
1.86:1 (total 21597)	89.33	91.36	85.53	92.17	84.16	94.99

TABLE IX
QUANTITATIVE EVALUATION OF THE EFFECT OF BILATERAL FILTERING. PRE INDICATES THAT THE BILATERAL FILTERING WAS USED AS PREPROCESSING. MED MEANS THAT THE BILATERAL FILTERING OPERATION WAS ADDED AFTER EVERY CONVOLUTION BLOCKS IN CNNS.

Methods	Results (%)			
	ACC	SEN	SPE	AUC
ResNet-18	85.06	88.08	79.42	91.79
ResNet-18/Pre	86.78	88.95	82.73	92.64
ResNet-18/Med	85.24	86.77	82.40	91.29
LoGo-Net	89.33	91.36	85.53	94.99
LoGo-Net/Pre	89.36	92.29	83.89	94.78
LoGo-Net/Med	88.90	91.41	84.22	94.81

Considering that the proportion of malignant nodules is less than 10% among the whole incidence of thyroid nodules [4] and the malignant/benign ratio of the THY-BM dataset around 2:1, we conducted further experiments based on five-fold cross-validation by controlling the total number of training images to be the same and used two malignant/benign ratios of 1:9 and 2:1. The results are listed in Table VIII. We can see that there is no apparent difference between the results of the two ratios, which indicates that the model trained with different proportions of the dataset does not cause an excessively biased diagnosis of malignancy.

F. Effectiveness of bilateral filtering

Bilateral filtering has good properties in effectively eliminating noise while maintaining relatively complete image structure [58]–[60]. Some studies have proposed that adding denoising as image preprocessing or adding bilateral filtering operation into intermediate layer in CNNs can improve the performance [61], [62]. These studies suggested us to conduct similar experiments since ultrasound images are also contain various types of noise. The results are shown in the Table IX.

Bilateral filter preprocessing has improved performance in ResNet-18, but the results of our LoGo-Net with such preprocessing are not significantly different from the previous results. This is because noise affects the learning of CNNs features to a certain extent, and the preprocessing operation plays a role in alleviating the effect of noise. However, guided by dual pathways, our LoGo-Net can obtain features more accurately, which makes the contribution of bilateral filtering unnoticeable in our method.

In addition, adding the bilateral filtering operation to the intermediate layer of CNNs does not improve the performance. In fact, the fundamental purpose of denoising is not to improve the visual effect but to assist the downstream tasks. Therefore, it is a valuable topic to consider the combination of denoising and classification/segmentation to establish a multi-task learning framework to promote different tasks [63].

TABLE X
PARAMETERS AND INFERENCE TIME OF DIFFERENT MODELS.

Methods	FLOPs(GB)	Parameters(MB)	FPS/GPU	FPS/CPU
Chi et al. [17]	1.50	5.60	54.45	11.70
Wang et al. [36]	6.48	54.31	19.46	4.65
Ma et al. [18]	4.53	193.40	104.64	16.73
Liu et al. [37]	3.16	177.90	94.19	12.20
Foo et al. [44]	14.49	16.48	188.50	8.53
Zhou et al. [43]	10.25	12.27	204.46	12.86
ResNet-18 [24]	1.82	11.18	172.79	24.17
LoGo-Net	3.17	13.69	112.45	5.70

G. Real-time testing

We assessed the floating-point operations (FLOPs), parameters, and frames per second (FPS) of different models, as shown in Table X. The experiment was conducted based on an Intel(R) Xeon(R) Gold 6230 CPU and an NVIDIA RTX 3090 GPU. In order to obtain the long-distance dependency representation of features, our TAD block introduces more complex self-attention calculations, which makes the model complexity increased compared with the baseline model. In general, our LoGo-Net introduces a small number of additional parameters and can ensure high real-time performance when using a single GPU or a single CPU, making it easier to migrate to edge computing platforms.

V. DISCUSSION AND CONCLUSION

In this work, our objective was to use ultrasound images to classify benign and malignant thyroid nodules. Inspired by the medical domain knowledge, we found that the basis of diagnosis comes from the doctor's abstract representation of imaging manifestation, which includes two parts: local tissue features and global anatomy features. Therefore, effectively establishing feature expression combining local and global features was the core of this research.

The proposed LoGo-Net was inspired by the human visual system, which divides the visual information processing into two paths with a hierarchical structure [25]. The information transmitted by the "What" pathway is related to target recognition, and the information transmitted by the "Where" pathway is related to target location. These two subsystems together constitute the perception model. To enable the information along the two channels to interact and promote each other, we designed the TAD block to decouple features into two components: the tissue component and the anatomy component. The

tissue component is used to capture local feature clues, and the anatomy component is used to capture global feature clues. After multi-task learning, the model obtained state-of-the-art performance.

To verify the practical clinical value of LoGo-Net, we built a large-scale dataset with a data volume that has not been reached in previous studies. The ultrasound images were acquired from different types of ultrasound equipment, so they can verify the universality of the proposed model. We visualized the saliency maps of the network to observe the AI's discrimination area when classifying benign and malignant thyroid nodules. The proposed method achieves excellent recognition of nodules with significant morphological differences, and the attended regions are more consistent with those of human experts, which increases the interpretability of deep learning to a certain extent.

In summary, we proposed an AI-based ultrasound image analysis method to diagnose benign and malignant thyroid nodules. Its excellent diagnostic ability gives it the potential to be used in the clinical work of doctors, including early cancer screening procedures in remote or resource-poor areas. In future studies, we will attempt to extend the proposed method to other similar ultrasound imaging tasks. Moreover, we will try to continue learning from the doctor's diagnosis model to fuse images with various modality data (e.g., patients' history codes) to make more accurate decisions in more complex clinical scenarios.

ACKNOWLEDGMENT

We thank LetPub for its linguistic assistance during the preparation of this manuscript.

REFERENCES

[1] G. H. Tan and H. Gharib, "Thyroid incidentalomas: management approaches to nonpalpable nodules discovered incidentally on thyroid imaging," *Ann. Intern. Med.*, vol. 126, no. 3, pp. 226–231, 1997.

[2] S. Ezzat, D. A. Sarti, D. R. Cain, and G. D. Braunstein, "Thyroid incidentalomas: prevalence by palpation and ultrasonography," *Arch. Intern. Med.*, vol. 154, no. 16, pp. 1838–1840, 1994.

[3] N. C. Institute, "Thyroid cancer screening," [EB/OL], August 2020, https://www.cancer.gov/types/thyroid/patient/thyroid-screening-pdq.

[4] E. Papini, R. Guglielmi, A. Bianchini, A. Crescenzi, S. Taccogna, F. Nardi *et al.*, "Risk of Malignancy in Nonpalpable Thyroid Nodules: Predictive Value of Ultrasound and Color-Doppler Features," *The Journal of Clinical Endocrinology & Metabolism*, vol. 87, no. 5, pp. 1941–1946, 05 2002.

[5] E. Koike, S. Noguchi, H. Yamashita, T. Murakami, A. Ohshima, H. Kawamoto *et al.*, "Ultrasonographic characteristics of thyroid nodules: prediction of malignancy," *Archives of surgery*, vol. 136, no. 3, pp. 334–337, 2001.

[6] J. P. Brito, M. R. Gionfriddo, A. Al Nofal, K. R. Boehmer, A. L. Leppin, C. Reading *et al.*, "The accuracy of thyroid nodule ultrasound to predict thyroid cancer: systematic review and meta-analysis," *The Journal of Clinical Endocrinology & Metabolism*, vol. 99, no. 4, pp. 1253–1263, 2014.

[7] E. S. Cibas and S. Z. Ali, "The 2017 bethesda system for reporting thyroid cytopathology," *Thyroid*, vol. 27, no. 11, pp. 1341–1346, 2017.

[8] E. Horvath, S. Majlis, R. Rossi, C. Franco, J. P. Niedmann, A. Castro *et al.*, "An ultrasonogram reporting system for thyroid nodules stratifying cancer risk for clinical management," *The Journal of Clinical Endocrinology & Metabolism*, vol. 94, no. 5, pp. 1748–1751, 2009.

[9] J.-Y. Park, H. J. Lee, H. W. Jang, H. K. Kim, J. H. Yi, W. Lee *et al.*, "A proposal for a thyroid imaging reporting and data system for ultrasound features of thyroid carcinoma," *Thyroid*, vol. 19, no. 11, pp. 1257–1264, 2009.

[10] S.-P. Cheng, J.-J. Lee, J.-L. Lin, S.-M. Chuang, M.-N. Chien, and C.-L. Liu, "Characterization of thyroid nodules using the proposed thyroid imaging reporting and data system (ti-rads)," *Head & neck*, vol. 35, no. 4, pp. 541–547, 2013.

[11] G. Russ, B. Royer, C. Bigorgne, A. Rouxel, M. Bienvenu-Perrard, L. Leenhardt *et al.*, "Prospective evaluation of thyroid imaging reporting and data system on 4550 nodules with and without elastography," *Eur. J. Endocrinol.*, vol. 168, no. 5, pp. 649–55, 2013.

[12] Y. Xie, Y. Xia, J. Zhang, Y. Song, D. Feng, M. Fulham *et al.*, "Knowledge-based collaborative deep learning for benign-malignant lung nodule classification on chest ct," *IEEE Trans. Med. Imaging*, vol. 38, no. 4, pp. 991–1004, 2019.

[13] Y. Qin, J. Wen, H. Zheng, X. Huang, J. Yang, N. Song *et al.*, "Varifocal-net: A chromosome classification approach using deep convolutional networks," *IEEE Trans. Med. Imaging*, vol. 38, no. 11, pp. 2569–2581, 2019.

[14] R. J. Gillies, P. E. Kinahan, and H. Hricak, "Radiomics: images are more than pictures, they are data," *Radiology*, vol. 278, no. 2, pp. 563–577, 2016.

[15] S. J. Nam, J. Yoo, H. S. Lee, E.-K. Kim, H. J. Moon, J. H. Yoon *et al.*, "Quantitative evaluation for differentiating malignant and benign thyroid nodules using histogram analysis of grayscale sonograms," *Journal of Ultrasound in Medicine*, vol. 35, no. 4, pp. 775–782, 2016.

[16] U. Raghavendra, U. R. Acharya, A. Gudigar, J. H. Tan, H. Fujita, Y. Hagiwara *et al.*, "Fusion of spatial gray level dependency and fractal texture features for the characterization of thyroid lesions," *Ultrasonics*, vol. 77, pp. 110–120, 2017.

[17] J. Chi, E. Walia, P. Babyn, J. Wang, G. Groot, and M. Eramian, "Thyroid nodule classification in ultrasound images by fine-tuning deep convolutional neural network," *J. Digit. Imaging*, vol. 30, no. 4, pp. 477–486, 2017.

[18] J. Ma, F. Wu, J. Zhu, D. Xu, and D. Kong, "A pre-trained convolutional neural network based method for thyroid nodule diagnosis," *Ultrasonics*, vol. 73, pp. 221–230, 2017.

[19] A. Zwanenburg, M. Vallières, M. A. Abdalah, H. J. Aerts, V. Andrearczyk, A. Apte *et al.*, "The image biomarker standardization initiative: standardized quantitative radiomics for high-throughput image-based phenotyping," *Radiology*, vol. 295, no. 2, pp. 328–338, 2020.

[20] X. Xie, J. Niu, X. Liu, Z. Chen, S. Tang, and S. Yu, "A survey on incorporating domain knowledge into deep learning for medical image analysis," *Medical Image Analysis*, vol. 69, p. 101985, 2021.

[21] X. He, Y. Deng, L. Fang, and Q. Peng, "Multi-modal retinal image classification with modality-specific attention network," *IEEE Trans. Med. Imaging*, pp. 1–1, 2021.

[22] J. Ding, H. D. Cheng, J. Huang, and Y. Zhang, "Multiple-instance learning with global and local features for thyroid ultrasound image classification," in *2014 7th International Conference on Biomedical Engineering and Informatics*, 2014, pp. 66–70.

[23] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural Computation*, vol. 1, no. 4, pp. 541–551, 1989.

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE CVPR*, June 2016.

[25] J. Norman, "Two visual systems and two theories of perception: An attempt to reconcile the constructivist and ecological approaches," *Behavioral and brain sciences*, vol. 25, no. 1, p. 73, 2002.

[26] J. M. Wolfe, M. L.-H. Võ, K. K. Evans, and M. R. Greene, "Visual search in scenes involves selective and nonselective pathways," *Trends in cognitive sciences*, vol. 15, no. 2, pp. 77–84, 2011.

[27] M. Himmelbach and H.-O. Karnath, "Dorsal and ventral stream interaction: contributions from optic ataxia," *Journal of Cognitive Neuroscience*, vol. 17, no. 4, pp. 632–640, 2005.

[28] L. Itti and C. Koch, "A saliency-based search mechanism for overt and covert shifts of visual attention." *Vision Research*, vol. 40, no. 10, pp. 1489–1506, 2000.

[29] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez *et al.*, "Attention is all you need," in *NIPS*, 2017, pp. 6000–6010.

[30] G. Song, F. Xue, and C. Zhang, "A model using texture features to differentiate the nature of thyroid nodules on sonography," *Journal of Ultrasound in Medicine*, vol. 34, no. 10, pp. 1753–1760, 2015.

[31] H. A. Nugroho, M. Rahmawaty, Y. Triyani, and I. Ardiyanto, "Texture analysis for classification of thyroid ultrasound images," in *2016 International Electronics Symposium (IES)*, 2016, pp. 476–480.

[32] A. Esteva, B. Kuprel, R. A. Novoa, J. M. Ko, S. M. Swetter, H. M. Blau *et al.*, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, 2017.

[33] B. Lei, S. Huang, H. Li, R. Li, C. Bian, Y.-H. Chou *et al.*, "Self-co-attention neural network for anatomy segmentation in whole breast ultrasound," *Med. Image Anal.*, vol. 64, p. 101753, 2020.

[34] Y. Zhou, H. Chen, Y. Li, Q. Liu, X. Xu, S. Wang *et al.*, "Multi-task learning for segmentation and classification of tumors in 3d automated breast ultrasound images," *Med. Image Anal.*, vol. 70, p. 101918, 2021.

[35] T. Liu, S. Xie, J. Yu, L. Niu, and W. Sun, "Classification of thyroid nodules in ultrasound images using deep model based transfer learning and hybrid features," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 919–923.

[36] L. Wang, L. Zhang, M. Zhu, X. Qi, and Z. Yi, "Automatic diagnosis for thyroid nodules in ultrasound images by deep neural networks," *Med. Image Anal.*, vol. 61, p. 101665, 2020.

[37] T. Liu, Q. Guo, C. Lian, X. Ren, S. Liang, J. Yu *et al.*, "Automated detection and classification of thyroid nodules in ultrasound images using clinical-knowledge-guided convolutional neural networks," *Med. Image Anal.*, vol. 58, p. 101555, 2019.

[38] C. Pei, F. Wu, L. Huang, and X. Zhuang, "Disentangle domain features for cross-modality cardiac image segmentation," *Med. Image Anal.*, vol. 71, p. 102078, 2021.

[39] A. Chartsias, G. Papanastasiou, C. Wang, S. Semple, D. E. Newby, R. Dharmakumar *et al.*, "Disentangle, align and fuse for multimodal and semi-supervised image segmentation," *IEEE Trans. Med. Imaging*, vol. 40, no. 3, pp. 781–792, 2021.

[40] Q. Meng, J. Matthew, V. A. Zimmer, A. Gomez, D. F. A. Lloyd, D. Rueckert *et al.*, "Mutual information-based disentangled neural networks for classifying unseen categories in different domains: Application to fetal ultrasound imaging," *IEEE Trans. Med. Imaging*, vol. 40, no. 2, pp. 722–734, 2021.

[41] Y. Tang, Y. Tang, Y. Zhu, J. Xiao, and R. M. Summers, "A disentangled generative model for disease decomposition in chest x-rays via normal image synthesis," *Med. Image Anal.*, vol. 67, p. 101839, 2021.

[42] P. Moeskops, J. M. Wolterink, B. H. M. van der Velden, K. G. A. Gilhuijs, T. Leiner, M. A. Viergever *et al.*, "Deep learning for multi-task medical image segmentation in multiple modalities," in *MICCAI*, 2016, pp. 478–486.

[43] Y. Zhou, H. Chen, Y. Li, Q. Liu, X. Xu, S. Wang, P.-T. Yap, and D. Shen, "Multi-task learning for segmentation and classification of tumors in 3d automated breast ultrasound images," *Medical Image Analysis*, vol. 70, p. 101918, 2021.

[44] A. Foo, W. Hsu, M. L. Lee, G. Lim, and T. Y. Wong, "Multi-task learning for diabetic retinopathy grading and lesion segmentation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 08, 2020, pp. 13 267–13 272.

[45] Z. Lin, S. Li, D. Ni, Y. Liao, H. Wen, J. Du *et al.*, "Multi-task learning for quality assessment of fetal head ultrasound images," *Medical Image Analysis*, vol. 58, p. 101548, 2019.

[46] S. Chen, G. Bortsova, A. García-Uceda Juárez, G. van Tulder, and M. de Bruijne, "Multi-task attention-based semi-supervised learning for medical image segmentation," in *MICCAI*, 2019, pp. 457–465.

[47] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE CVPR*, 2018, pp. 7794–7803.

[48] M. Yin, Z. Yao, Y. Cao, X. Li, Z. Zhang, S. Lin *et al.*, "Disentangled non-local neural networks," in *ECCV*. Springer International Publishing, 2020, pp. 191–207.

[49] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*, 2015, pp. 234–241.

[50] J. H. Lee, J. H. Baek, J. H. Kim, W. H. Shim, S. R. Chung, Y. J. Choi *et al.*, "Deep learning–based computer-aided diagnosis system for localization and diagnosis of metastatic lymph nodes on ultrasound: A pilot study," *Thyroid*, vol. 28, no. 10, pp. 1332–1338, 2018.

[51] C. Li, C. Xu, C. Gui, and M. D. Fox, "Distance regularized level set evolution and its application to image segmentation," *IEEE Transactions on Image Processing*, vol. 19, no. 12, pp. 3243–3254, 2010.

[52] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE CVPR*, June 2016.

[53] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. Torr, "Res2net: A new multi-scale backbone architecture," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 2, pp. 652–662, 2021.

[54] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE CVPR*, July 2017.

[55] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov *et al.*, "Going deeper with convolutions," in *Proc. IEEE CVPR*, June 2015.

[56] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *arXiv preprint: 1312.6034*, 2014.

[57] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, and D. Ren, "Distance-iou loss: Faster and better learning for bounding box regression," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 12 993–13 000.

[58] R. Gadde, V. Jampani, M. Kiefel, D. Kappler, and P. Gehler, "Superpixel convolutional networks using bilateral inceptions," in *ECCV*, 2016, pp. 597–613.

[59] S. Ghosh and K. N. Chaudhury, "Fast and high-quality bilateral filtering using gauss-chebyshev approximation," in *SPCOM*, 2016, pp. 1–5.

[60] ——, "Color bilateral filtering using stratified fourier sampling," in *GlobalSIP*, 2018, pp. 26–30.

[61] V. S. Unni, S. Ghosh, and K. N. Chaudhury, "Linearized admm and fast nonlocal denoising for efficient plug-and-play restoration," in *2018 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, 2018, pp. 11–15.

[62] S. Ghosh, A. K. Mandal, and K. N. Chaudhury, "Pruned non-local means," *IET Image Processing*, vol. 11, no. 5, pp. 317–323, 2017.

[63] M. Eslami, S. Tabarestani, and M. Adjouadi, "Joint low dose ct denoising and kidney segmentation," in *ISBI Workshops*, 2020, pp. 1–4.